JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1990, 23, 323–331 NUMBER 3 (FALL 1990)

THE REPRESENTATIVENESS OF OBSERVATIONAL SAMPLES OF DIFFERENT DURATIONS

OLIVER C. MUDFORD AND IVAN L. BEALE
UNIVERSITY OF AUCKLAND, NEW ZEALAND

AND

NIRBHAY N. SINGH
VIRGINIA COMMONWEALTH UNIVERSITY

The representativeness of behavioral observation samples with durations of less than the whole time of interest was investigated. A real-time recording system was developed to quantify the behavior of 5 profoundly mentally retarded physically handicapped adult students in an institutional training setting. Behavior was observed using six mutually exclusive and exhaustive categories during 2.5-hr observation sessions. Sample observation sessions with durations ranging from 15 to 135 min were computer simulated from the whole-session (150-min) records. It was found that the representativeness of these samples, when compared to whole-session records, was a function of the relative duration of the behavioral categories and of sample duration. The occurrence of relatively high-duration behaviors (lasting for more than 50% of the session) was estimated to within 20% error by samples of less than 60 min, but low-duration behaviors (1 to 3% of the session) were inadequately quantified even from 135-min samples. Increasing irregularity of bouts of behavior in the low-duration behaviors is suggested as the cause of the functions obtained. Implications of the findings for applied behavior analysis are discussed, with the recommendation that the adequacy of observational session durations be empirically assessed routinely.

DESCRIPTORS: behavioral assessment, behavioral observation, measurement error, time sample, mentally retarded adults

Assessment of performance by direct observational methods is one of the distinguishing characteristics of applied behavior analysis. In contrast to psychometric measurement devices, there is generally no need to argue that what is being measured represents the behaviors of interest. However, it has long been recognized that the quality of observational measurement needs to be thoroughly investigated. At issue are the effects of superimposing sampling methods on ongoing streams of behavior when the components cannot, for practical reasons, be measured continuously, with known accuracy, in all settings, and through the whole time of interest.

The effects of noncontinuous measurement by interval and time sampling methods have received considerable attention (e.g., Harrop & Daniels, 1986; Repp, Roberts, Slack, Repp, & Berkler, 1976). The accuracy of measurement, whether estimated directly or through inference from interobserver agreement, has been extensively studied (e.g., Boykin & Nelson, 1981). The differential results of behavior assessment across settings have been noted and recognized by behavior analysts using multiple baseline across settings designs (e.g., Odom, Hoyson, Jamieson, & Strain, 1985). Scant attention, however, has been given to the effect of observation session duration within a setting, although the issue has often been raised (e.g., Altmann, 1974; Goldfried, 1983; Hartmann, 1984; Wildman & Erickson, 1977).

Some studies relevant to the question of duration of observation have been performed by educational

Some of the data were presented to the Applied Behavior Analysis symposium at the conference of the New Zealand Psychological Society at Dunedin, New Zealand in August 1986.

We thank Jane Penney (Templeton Hospital) for assistance in the conduct of the study and Mike Owen (Auckland University) for producing the figures.

Reprint requests should be sent to Ivan L. Beale, Department of Psychology, University of Auckland, Auckland, New Zealand.

researchers (e.g., Karweit & Slavin, 1982; Rowley, 1978). Their approach has been to use psychometric concepts of generalizability to assess the stability of classroom behavior across samples, rather than the degree to which the sample represents the whole. This approach may be invalidated by unwarranted assumptions about the distributions of the behaviors sampled and by the failure to test for trend (Rogosa, Floden, & Willett, 1984). The generality of their findings is further restricted by the measurement of only frequency data using momentary time sampling.

The validity of observational samples with respect to longer time periods has been investigated in assessing the behaviors of psychiatric patients (Alevizos, DeRisi, Liberman, Eckman, & Callahan, 1978). This study evaluated the representativeness of data obtained from two 15-s observations per day against a criterion measure obtained from 15 such observations over 12 hr. However, the representativeness of the criterion data was not assessed against the whole time of interest (i.e., the waking hours of the day). Further, the validity of the recording method was not assessed against a continuous record, so possible invalidity due to recording confounds interpretation of their results.

The problem of selecting a duration for observation sessions has yet to be solved, although some recommendations have been made. For instance, Bijou, Peterson, Harris, Allen, and Johnston (1969) and Kazdin (1984) have suggested 1-hr observation periods. Only Johnston and Pennypacker (1980) appear to have suggested that all responses may have to be observed, at least temporarily, to assess empirically the representativeness of samples smaller than the whole time of interest. This recommendation can be seen as analogous to that of Sanson-Fisher, Poole, and Dunn (1980) that appropriate interval lengths for interval recording ought to be empirically determined by simulated sampling of real-time records of behavior.

The present study used trained observers to take real-time continuous records of behavior for the whole time of interest. These records served as criteria for comparison with computer-simulated sample sessions of varying duration drawn from the whole-session records. Subsequently, sample sessions of adequate length were subjected to computer-simulated momentary time sampling to provide an example of the compounding of errors produced by sample length and interobservation interval length.

METHOD

Subjects and Setting

Five profoundly mentally retarded adults were selected from a class of 10 attending a training program in a residential facility. Selection was based only on regularity of attendance. Ages ranged between 26 and 35 years, and length of stay in the institution varied from 6 to 28 years. All subjects had impaired mobility and used wheelchairs or other assistance to move.

The 3 training staff were teachers of people with severe and profound handicaps. Observations were made in the training area (14 m by 7 m) throughout morning and afternoon training sessions (8:30 to 11:00 a.m. and 1:00 to 3:30 p.m.). Residents were generally provided with training materials more suited to an educational curriculum than a functional curriculum (Reid et al., 1985). Morning and afternoon sessions differed in that residents sat at individual tables in the morning and around a large table in the afternoon. Staff were not aware of the purpose of the study.

Apparatus

A portable IBM personal computer (PC) was programmed in BASIC to record real-time observational data.

Observation Categories

An exhaustive and mutually exclusive set of categories was developed to describe subjects' behaviors (Sackett, 1978). Six categories were selected that had face validity for assessing the subjects' activities: social interaction with peer (SP); social interaction with staff (SS); handling materials provided (HM); self-propelled movement (SM); inappropriate behaviors, including stereotyped and self-injurious behaviors (I); and passive (P). (De-

tailed definitions may be obtained from the authors.)

Mutual exclusivity was obtained through the use of a priority coding system. The categories have been listed in order of priority. In practice, the effect of this system was to make the categories SM, HM, and P independent of staff assistance or social interaction. It had been observed before formal observation that inappropriate behavior was never concurrent with behavior categories higher in priority. A single-digit code was assigned to each category for input to the PC.

Observers

Five pairs of undergraduate students (who were enrolled in a third-year course in applied behavior analysis) and the first author acted as observers. Each pair of observers was assigned to 1 subject. The observers were trained in the training room and from videotapes until interobserver agreement, assessed by kappa, exceeded 0.75 in two 10-min sessions in the training room (Cohen, 1960; Hollenbeck, 1978).[1] Student observers were not informed of the purpose of the study.

Observation Procedure

Each subject was observed for one entire training day (i.e., the 5 hr spent in the training area, divided into two 2.5-hr sessions). Thus, in total, there were 10 150-min sessions recorded. Neither staff nor subject reactivity to observation was apparent. The training staff had been informed that the behaviors of individual staff members were not being assessed.

At session onset, the primary observer began entering behavior codes on the numeric keypad to the right of the PC keyboard. A printed list of codes was available during observations. Whenever the subject's behavior changed to that defined by a different category the observer entered a new code. Because categories were mutually exclusive, only the time of the start of a bout of behavior was stored in the PC's solid-state memory along with the code entered. These raw data were filed on disk at the end of the observation session for later analysis. Alphanumeric codes, which had been entered, were displayed on the PC screen to provide a visual check. When possible, primary observers alternated each half hour as a safeguard against fatigue. If, during observations, the subject became obscured from the primary observer, another observer followed the subject and hand-signaled changes in code.

Interobserver Agreement

Reliability of observations was assessed by comparing the records of two observers recording simultaneously. The second observer sat to the left of the primary observer and used letter codes to represent behavioral categories. Neither observer was informed of the codes used by the other, although the senior author/observer sometimes acted in either capacity. Observers were asked not to discuss coding during observations and were not heard to do so. Agreement checks were immediately terminated if the observers' view of the subject was obscured.

Interobserver agreement assessments were spaced throughout sessions and occupied between 26% and 34% of the training day for each subject. As the measure for agreement between observers, kappa was computed using the second-by-second algorithm detailed by Hollenbeck (1978). In 29 of the 30 checks, kappa exceeded the criterion value of 0.75. The mean value of kappa across observations was 0.89. No feedback was provided to observers on their levels of agreement.

RESULTS

The Whole-Session Records

The primary observer's record of each code was taken as a whole-session record. There were 60

[1] Kappa was used to quantify interobserver agreement because it appeared to be an acceptable coefficient of agreement for continuous observational records at the time of data collection (1985) (e.g., Sanson-Fisher et al., 1980). In hindsight, modifications of the well-known percentage agreement formulae may be seen as more appropriate indices of agreement between observers (MacLean, Tapp, & Johnson, 1985; Repp, Harman, Felce, Van Acker, & Karsh, 1989). Kappa by Hollenbeck's method and percentage agreement cannot be formally (i.e., mathematically) related.

such records: six codes and two records per subject. The absolute duration of behavior in sessions was computed by cumulating the time differences between onsets and offsets of a code. Relative durations of the code in each record were computed by transforming the absolute duration into the percentage of the 2.5-hr session. Across subjects and sessions the average percentage of time taken up with social interactions with peers was 3.0% (range, <0.1% to 11.1%); for social interaction with staff, 15.9% (8.8% to 30.8%); for handling materials, 28.8% (1.7% to 56.6%); for self-propelled movement, 5.9% (0.1% to 13.8%); for inappropriate behavior, 0.4% (0.0% to 1.1%); and for passive behavior, 46.0% (11.1% to 85.5%). Thirteen records having a relative duration of <0.7% were not analyzed further because they were considered appropriate only for event recording. These included eight of the 10 records for inappropriate behavior.

To summarize the basic parameters of behavior other than relative duration, the records were grouped according to relative duration; however the 10 records for SS were excluded because it was suspected that these were qualitatively different from the residents' records. For behaviors occurring in more than 50% of a record, the mean absolute frequency was 62, the mean absolute duration of a bout was 109 s (maximum 1,825 s), and the average interbout time (IBT) was 49 s (maximum 895 s). For behaviors occurring for between 10% and 25% of sessions, the mean frequency was 36, the mean duration was 32 s (maximum 325 s), and the mean IBT was 203 s (maximum 4,355 s). For the lowest relative duration group warranting further consideration (relative duration of 0.7% to 2.9%), the mean frequency was 11, mean duration was 17 s (maximum 175 s), and mean IBT was 509 s (maximum 4,785 s). This summary has not included behaviors that occurred between 3% and 10%, nor those between 25% and 50%, nor SS behaviors, but the trends in data generally were consistent with those data reported, showing that relative duration, frequency, and mean duration increased together while IBT decreased.

Sample Sessions

The whole-session records were sampled by computer to permit inspection of the relation between sample sessions of various durations and the whole-session (150-min) records. The duration of the sample sessions ranged from 15 to 135 min, increasing from the lower figure by 15-min increments. At each of the nine sample session durations, three types of systematic samples were taken: centered on the midpoint of the whole session; beginning at the start of the whole session; and ending at the termination of the whole session. Subsequently, for five sample durations (15, 45, 75, 105, and 135 min) five random starting points were generated. The randomness was constrained by the sample duration.

Comparisons Between Sample Sessions and the Whole-Session Records

The relative duration of each code per sample was calculated as previously described for comparison with the whole-session relative duration. A percentage similarity statistic was computed by dividing the smaller of each pair by the larger and multiplying by 100. The resulting values were subtracted from 100 to yield a percentage difference score, in which zero indicates complete agreement and larger values indicate lesser agreement. When a relative duration value of zero was obtained for a sample, the resulting percentage difference score was 100.

For both systematic and random samples, percentage difference scores were grouped according to relative duration in the whole sessions (as above) and the mean percentage difference calculated. Again the records for SS were treated separately. Only data from samples centered on the midpoint of a session are included here (in Figure 1) because values obtained from all three systematic starting points were similar.

The functions plotted in Figure 1 show clear trends, with difference scores decreasing with increasing sample duration. There is also a clear effect of the relative duration parameter. The functions

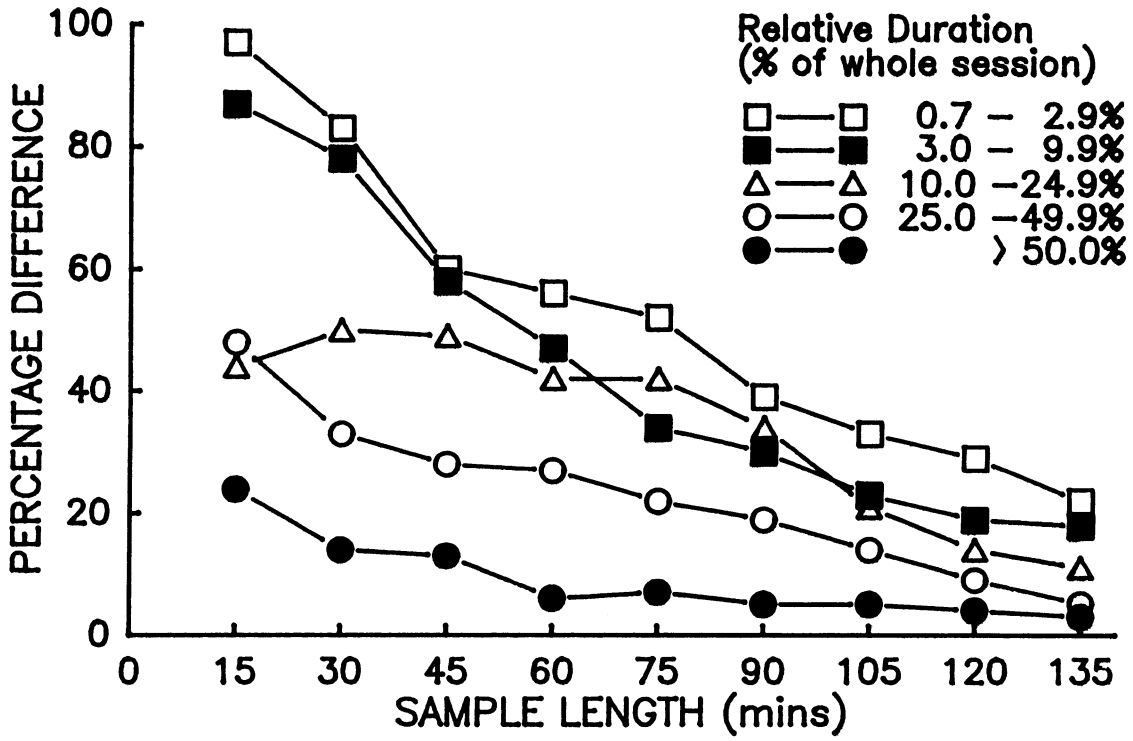

Figure 1. Percentage difference between relative duration from samples of increasing length and from the whole (150 min) sessions for subjects' codes grouped by whole-session relative duration. Samples were centered on the midpoint of the session.

tend to be vertically separated, percentage difference increasing with relative duration. There appears to be an interaction between relative duration and sample length, because values for the five relative duration groups differ more at short than at long sample durations.

Our next step was to explore the generality of the obtained effects of duration and session length across measures of difference. The random sample data were reanalyzed using a percentage error measure of correspondence between sample and whole-session records (Rojahn & Kanoy, 1985). This was computed by subtracting the sample value for relative duration from the whole-session value, dividing by the whole-session value, and multiplying by 100. Values of percentage error can range from +100% when the relative duration in the sample is zero to very large negative values when the sample provides a gross overestimate. The obtained values for percentage difference and percentage error are plotted in Figure 2. To facilitate comparison between the measures, the sign of the error score was made positive before averaging. Thus, only the magnitude, and not the direction, of the error is considered. Clearly, considering Figure 2, percentage error shows the same effects of relative duration and sample length revealed by the percentage dif-

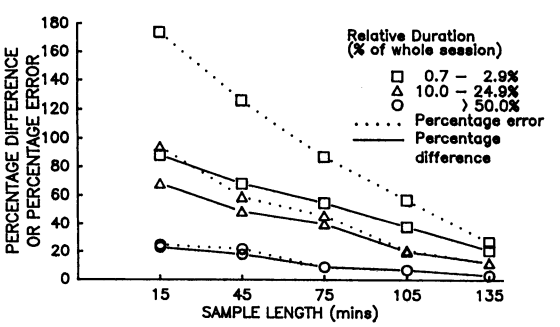

Figure 2. A comparison of the percentage difference and percentage error measures for groups of subjects' codes with low, moderate, and high relative durations.

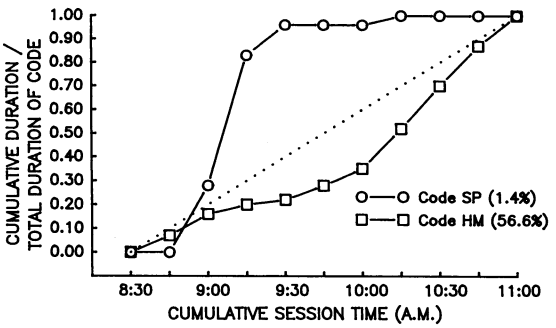

Figure 3. Cumulative duration/total duration for a high relative duration code (HM) and a low relative duration code (SP) from a morning session (8:30 to 11:00 a.m.). Proportion of code duration recorded is plotted against cumulative session time.

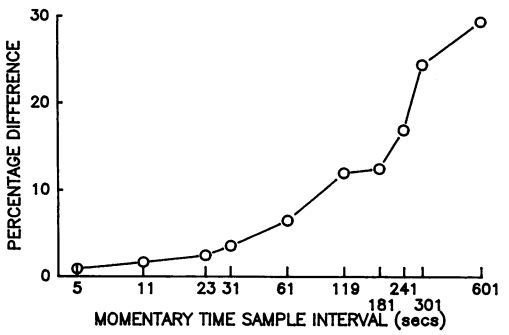

Figure 4. Percentage difference between real relative duration and relative duration derived from simulated momentary time samples of increasing interobservation intervals (seconds, logarithmic scale). Passive behavior was measured from three randomly selected hours in each session. Each data point represents the mean of 30 measures.

ference scores. However, both main effects and the interaction are magnified by the error measure.

In order that an explanation of the effects obtained may be offered, an analysis was made of the distributions of high and low relative duration behaviors in one observational session. The absolute duration of a high relative duration code (HM; relative duration = 56.6%) and of a low relative duration code (SP; 1.4%) was calculated in 15-min blocks throughout the session. Absolute durations were cumulated across successive blocks and divided by the total absolute duration of codes for that session. This produced a measure of cumulative duration as a proportion of total duration, which was plotted against cumulative session duration in Figure 3.

The dotted diagonal line in Figure 3 represents the theoretical cumulative function for a code, the occurrence of which is uniformly distributed across successive 15-min blocks of session time. Samples as small as a single block would accurately represent the relative duration of that code over the whole session. The low duration code occurs mainly between 8:45 and 9:30 a.m. Samples taken within that period would greatly overestimate the whole-session duration, and samples taken outside that period would grossly underestimate it. The high duration code is more uniformly distributed, however, and generally remains closer to the diagonal. Samples of the whole-session record of this code would generally be more representative than would samples of the low relative duration code.

A further analysis was conducted to explore the implications of these results for sampling methods not involving real-time recording. Momentary time sampling (MTS) was chosen because it is considered the least biased method for estimating duration (e.g., Harrop & Daniels, 1986) and is probably less demanding on the observer than most alternatives. Three randomly selected 60-min segments from some whole-session records were sampled by a computer program simulating MTS at intervals from 5 s to 601 s. Percentage difference between obtained MTS data and the real-time data for duration within the segment was computed for relative durations of the code for passive behavior, which had the highest average relative duration across sessions and subjects at 46%. The resulting function is plotted in Figure 4 and shows that difference scores increase with increasing intervals between observations. If up to 25% difference is taken as an acceptable level of representativeness, it can be seen that an observation every 301 s is sufficient for assessment of the duration of the behavior in that hour. If less than 20% difference is desired, this can be achieved by one observation every 241 s.

DISCUSSION

This study investigated the representativeness of data obtained from observational samples with durations shorter than the whole time of interest (2.5

hr). It was found that increasing sample duration produced reduced error or difference when relative durations obtained from the samples were compared with relative durations across the whole time of interest. Generally, at any given sample length, behaviors of greater relative duration were sampled in a more representative fashion than those of smaller relative duration. Data summarizing the basic parameters of the behaviors suggest that these results could be explained by examining the distribution of behaviors across the whole times of interest. Increased relative duration was accompanied by increased frequency of bouts of the behavior, increased average duration of bouts, and decreased average IBTs. Further, obtained maximum values of the parameters exceeded the means by a factor of 10 to 20.

Such results might be predicted by rational analysis of the effects of sampling when events are irregular in their distribution across time. Successful prediction could also have been achieved from study of the results of analogous studies investigating interval or time sampling within observation sessions. For example, Green and Alverson (1978) determined that bias in recording at a given interval length was related to mean duration of behavior and mean IBT. In other words, with the obtained uneven distributions (e.g., in Figure 3) the obtained effects on representativeness were to be expected. Without prior data, however, the actual distributions of behaviors through the whole time of interest cannot be predicted, nor can appropriate values for parameters be chosen for computer-generated pseudobehaviors (e.g., Green & Alverson, 1978; Rojahn & Kanoy, 1985).

Less predictable are the findings regarding the *absolute* degree of error in the sample sessions of shorter duration. If, for example, 20% error is taken as the maximum acceptable, samples of at least 105-min duration were required for behaviors occurring only for 10% to 25% of the whole session. On the other hand, samples of only 30 min were adequately representative for behaviors taking up over 50% of the session. Thus, there is no support for the recommendation of a standard 60-min observation session (Bijou et al., 1969; Kazdin, 1984), even when the total time of interest is as little as 2.5 hr. The alternative recommendation of Johnston and Pennypacker (1980), that adequate observation session length ought to be empirically determined through exhaustive observation, has been strengthened. This parallels the findings of Sanson-Fisher et al. (1980) concerning the selection of an appropriate interval size for partial interval recording.

There are some limited cases to which this general recommendation does not apply (e.g., when the regularity of behavior can be known a priori). For example, in observations of a teacher or other behavior change agent performing according to predetermined schedules of prompting and reinforcement with unvarying durations and IBTs, observation of as little as one cycle of events may be representative of the whole series.

When speculating on the generality of the levels of absolute error, the characteristics of the present subjects, settings, and measurement system need to be considered. These observations were of nonambulatory profoundly retarded adults in a training setting that could best be described as archaic (Reid et al., 1985). If the training staff had been teaching their clients chronological age-appropriate functional skills such as self-propelled movement and social interactions with peers, the parameters of behaviors may have been quite different. However, baseline settings and levels of behaviors such as those described may not be infrequent. The use of mutually exclusive categories of behaviors in the present study may suggest a source of lack of generality, in that behaviors lower in priority were recorded only if higher priority behaviors were not co-occurring. This could result in the underestimation of the relative durations of the lower priority behaviors, with the concurrent effects of increasing IBTs, reducing bout durations, and reducing representativeness of small samples. Although no data were collected to counteract this criticism, it was informally observed that the effects of priority coding were as anticipated and described in the method section. In other words, the face validity of the recording system was not reduced by the method of formal observations.

An aspect of the measurement method used that may have implications for both internal and ex-

ternal validity of the study is the interobserver agreement assessment procedure. The observers were seated side by side, which strongly suggests a lack of independence (Kazdin, 1977). Even though they used different codes for behaviors and the behaviors were explicitly defined, the high levels of agreement obtained (as estimated by kappa) could have been due to observers cuing one another by key pressing rather than to the relatively easy job they had distinguishing behavior changes. In retrospect, it would have been preferable to produce an accurate criterion record of behavior as the whole-session record to be sampled (Johnston & Pennypacker, 1980).

The problem of error produced by the method of sampling within sessions is illustrated in Figure 4. In that case a sample-length error of up to about 20% was present before MTS was imposed on the session. If an acceptable MTS error of 20% was also present, these two errors compound to produce a 44% error if the sign of the error was the same. That magnitude would probably not be acceptable. Further compounding of error may be produced by observer error although, unless an accurate criterion record was produced for comparison, the result could not be combined mathematically with the other sources of error.

In summary, this study shows that, except in some special cases, the representativeness of observation sessions with respect to the whole time of interest should be empirically assessed. The absolute levels of error in sample sessions will differ across subjects, settings, and behavioral recording systems. Consideration should be given to the compounding effects of error produced by observers, method, and session duration.

REFERENCES

Alevizos, P., DeRisi, W., Liberman, R., Eckman, T., & Callahan, E. (1978). The behavior observation instrument: A method for program evaluation. *Journal of Applied Behavior Analysis,* **11,** 243–257.

Altmann, J. (1974). Observational study of behaviour: Sampling methods. *Behaviour,* **49,** 227–267.

Bijou, S. W., Peterson, R. F., Harris, F. K., Allen, E., & Johnston, M. S. (1969). Methodology for experimental studies of young children in natural settings. *Psychological Record,* **19,** 143–150.

Boykin, R. A., & Nelson, R. O. (1981). The effects of instructions and calculation procedures on observers' accuracy, agreement, and calculation correctness. *Journal of Applied Behavior Analysis,* **14,** 479–489.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* **20,** 37–46.

Goldfried, M. R. (1983). Behavioral assessment. In I. B. Weiner (Ed.), *Clinical methods in psychology* (2nd ed., pp. 233–281). New York: Wiley.

Green, S. B., & Alverson, L. G. (1978). A comparison of indirect measures for long duration behaviors. *Journal of Applied Behavior Analysis,* **11,** 530.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis,* **19,** 73–77.

Hartmann, D. P. (1984). Assessment strategies. In D. H. Barlow & M. Hersen (Eds.), *Single case experimental design: Strategies for studying behavior change* (2nd ed., pp. 107–139). New York: Pergamon.

Hollenbeck, A. R. (1978). Problems of reliability in observational research. In G. P. Sackett (Ed.), *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 79–98). Baltimore, MD: University Park Press.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research.* Hillsdale, NJ: Erlbaum.

Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology,* **74,** 844–851.

Kazdin, A. E. (1977). Artifact, bias, and complexity: The ABCs of reliability. *Journal of Applied Behavior Analysis,* **10,** 141–150.

Kazdin, A. E. (1984). *Behavior modification in applied settings* (3rd ed.). Homewood, IL: Dorsey.

MacLean, W. E., Tapp, J. T., & Johnson, W. L. (1985). Alternate methods and software for calculating interobserver agreement for continuous observation data. *Journal of Psychopathology and Behavioral Assessment,* **7,** 65–73.

Odom, S. L., Hoyson, M., Jamieson, B., & Strain, P. S. (1985). Increasing handicapped preschoolers' peer social interactions: Cross-setting and component analysis. *Journal of Applied Behavior Analysis,* **18,** 3–16.

Reid, D. H., Parsons, M. B., McCarn, J. E., Green, C. W., Phillips, J. F., & Schepis, M. M. (1985). Providing a more appropriate education for severely handicapped persons: Increasing and validating functional classroom tasks. *Journal of Applied Behavior Analysis,* **18,** 289–301.

Repp, A. C., Harman, M. L., Felce, D., Van Acker, R., & Karsh, K. G. (1989). Conducting behavioral assessments on computer-collected data. *Behavioral Assessment,* **11,** 249–268.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis,* **9,** 501–508.

Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing

the stability of teacher behavior. *Journal of Educational Psychology,* **76,** 1000–1027.

Rojahn, J., & Kanoy, R. C. (1985). Toward an empirically based parameter selection for time-sampling systems. *Journal of Psychopathology and Behavioral Assessment,* **7,** 99–120.

Rowley, G. L. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. *Journal of Educational Measurement,* **15,** 165–180.

Sackett, G. P. (1978). Measurement in observational research. In G. P. Sackett (Ed.), *Observing behavior: Vol. 2. Data collection and analysis methods* (pp. 25–43). Baltimore, MD: University Park Press.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis,* **13,** 493–500.

Wildman, B. G., & Erickson, M. T. (1977). Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment* (pp. 255–273). New York: Brunner/Mazel.

Received October 19, 1988
Initial editorial decision February 23, 1989
Revisions received November 3, 1989; December 15, 1989
Final acceptance April 18, 1990
Action Editor, Terry J. Page